

Tvorba slovníku

PLIN035: Počítačová lexikografie

František Kovařík

Opakování

První „slovníky“ ~ hliněné tabulky.

→ Knihy. (Psané, tištěné.)

→ → Digitalizace.

60. léta

Použití počítačů. (Papír → databáze.)

Automatická kontrola cross-referencí.

Brown Corpus.

70. léta

1978:

*Longman Dictionary
of Contemporary English*

80. léta

1980: COBUILD (!!!)

– Korpus současných textů.

(Bank of English)

80. léta

1987:

Collins COBUILD English

Language Dictionary

1987: COBUILD dict

„If a person, animal, or other living thing is killed, something or someone causes them to die.“

TEI

1987: Text Encoding Initiative

1990: TEI Guidelines

=> XML, DTD

90. léta

Vývoj prvních DWS

~ Dictionary Writing Systems.

3 revoluce v tvorbě slovníků

- 1) Využití korpusů.**
- 2) Přejechod k digitálu.**
- 3) Automatizace.**

Budoucnost?

The Oxford English Dictionary, 3rd Edition:

2000–2037?

Akademický slovník současné češtiny:

2005, 2012–2016–???

Dnešní téma

Jak dělat slovník.

Jak nedělat slovník.

Preskripce a deskripce

„Most British people say ‘different to’ and most Americans say ‘different than’, and they are all wrong because the ONLY correct form is ‘different from’.“

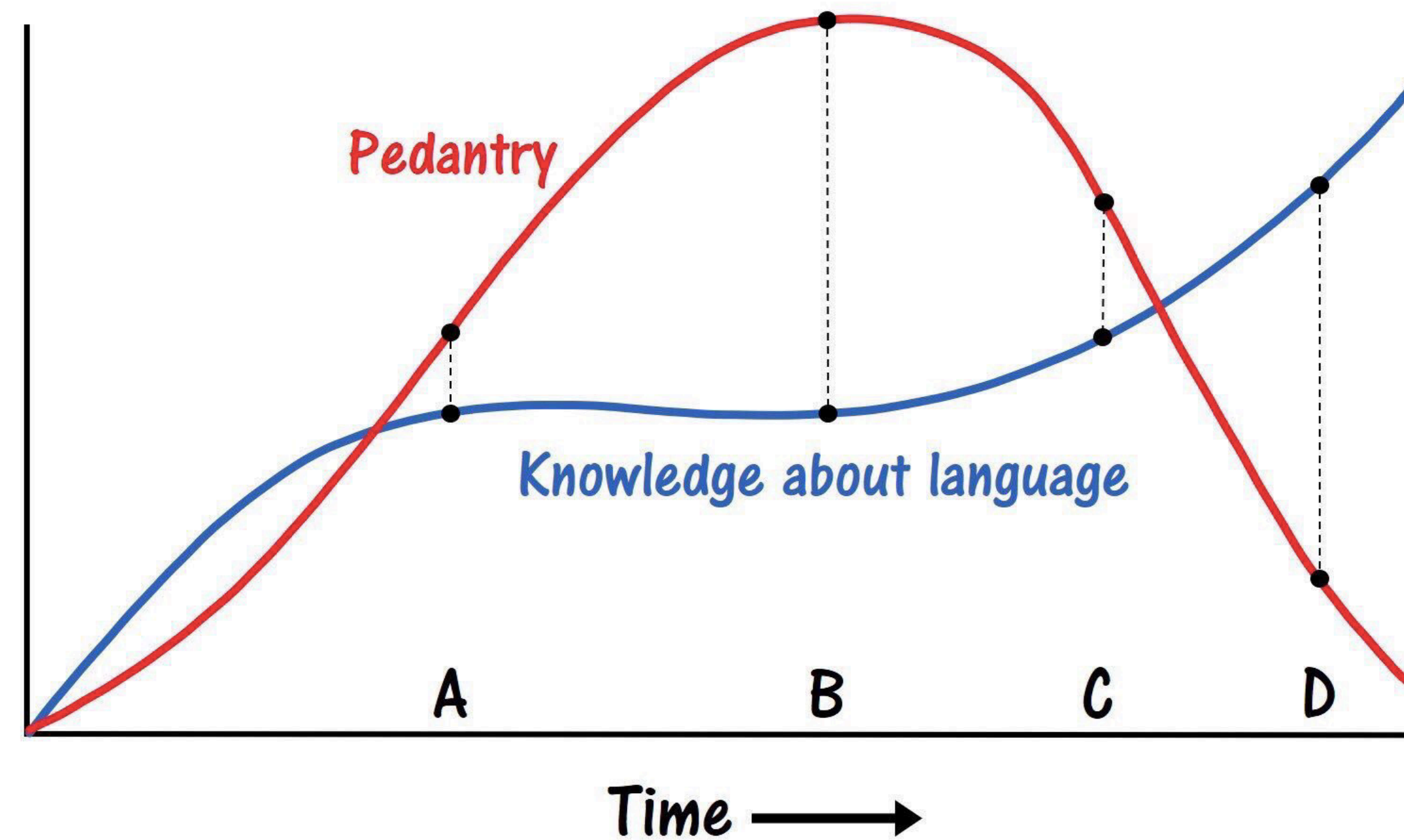
– The Queen’s English Society

Preskripce a deskripce

„At a conservative estimate, it can be said that some 90% of the English mother-tongue people who use this word [different] use it wrongly. Even if that were to rise to 100%, their use would still be wrong because there are some very basic and logical reasons why that is so.“

A graph and explanation of linguistic knowledge vs linguistic pedantry

@robdrummond



- A.** Danger point. Increased knowledge about language has led to a surge in linguistic pedantry, resulting in an imbalance.
- B.** Crisis point. Full on pedant. False confidence in good linguistic knowledge has led to pedantic half-truths outweighing facts.
- C.** The road to recovery. Taking the time to understand how language actually works has led to a more tolerant outlook.
- D.** A sensible place to be. The more you know about language, the more you see the nonsense of linguistic pedantry. Be aware that at point (D) you may be called upon to challenge people at point (B). This can be difficult, as people at point (B) often have the weight of public opinion behind them. Be patient, and encourage them to move towards point (C).

Preskripce a šamanismus

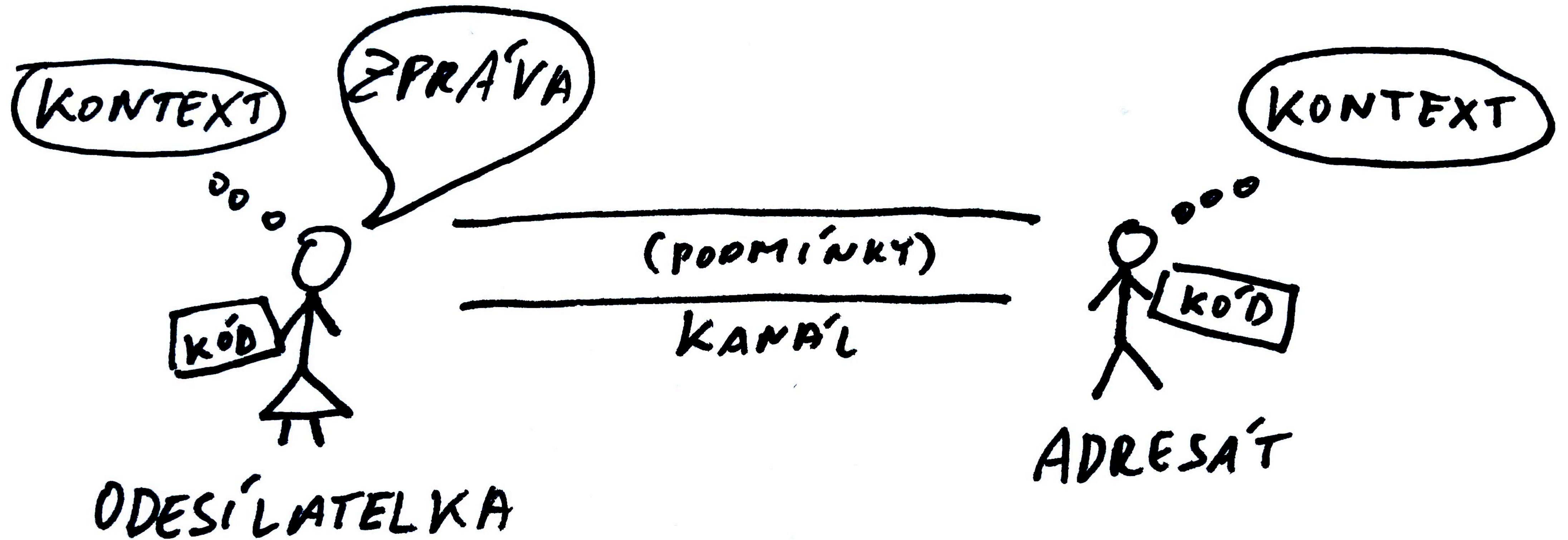


K čemu je preskripce?

K čemu je preskripce?

K porozumění.

Diagram komunikace (zjednodušený)

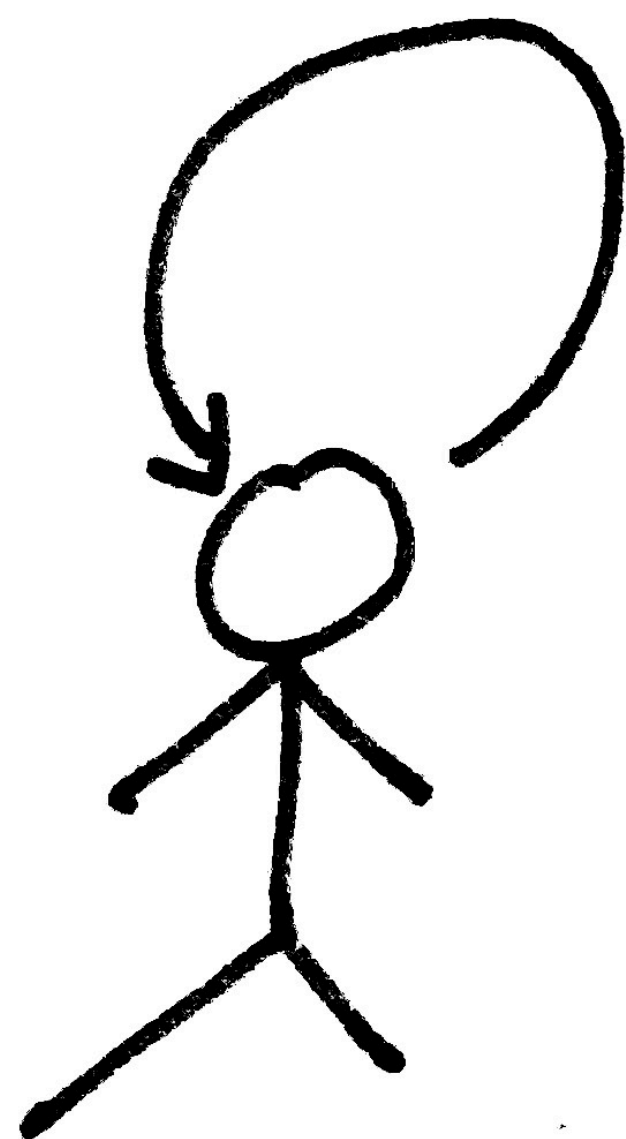


Meaning is use

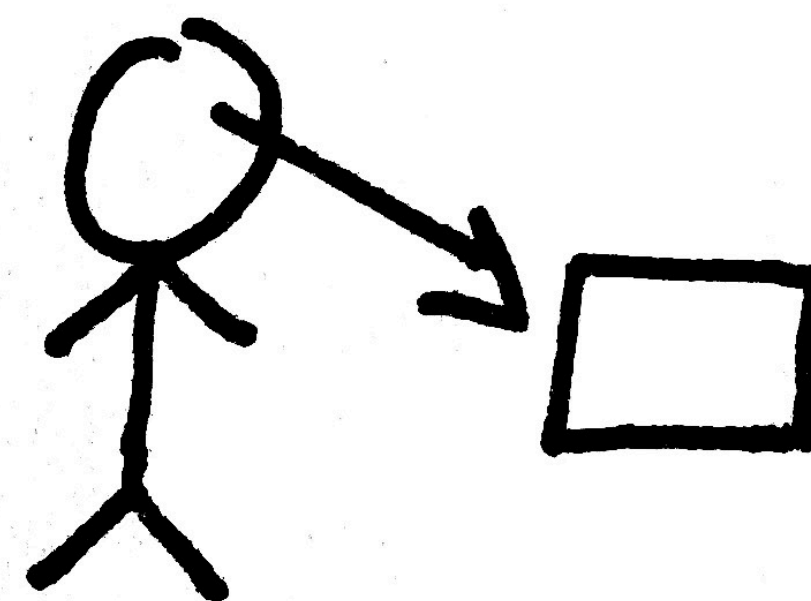
*„The meaning of a word is its use in the
language.“*

– Ludwig Wittgenstein, 1953

Introspekce vs. vnější pozorování



Introspekce



Vnější
pozorování

Deskripce

1857: Základní princip OED.

=> Evidence-based lexicography.

Evidence-based

Crowdsourcing.

Výstřižky od dobrovolníků.

→ Kolekce citací.

Excerpta



b u ř t m. /zde adv./

to by mně nebylo buřt

jedno, lhostejné

Jezdina poznamenal: - To se ti nedivím, brachu, to by
mně teda taky nebylo buřt...

1955(2.) J. V. Pleva: Eskortá

19, 9

/Pleva/

sranda f

vyládati ~ y

Havří vyláde své srandy a hří smolekú vápnou - ...

1949 Majerová-Lacina
Havří. anekdoty

82, 12

Detail hesla

Bleskový filtr

Heslo	<input type="text" value="myš"/>	Slovní druh	<input type="text" value="podst ž"/>
Podheslí	<input type="text" value="myš (polohovací to zařízením)"/>		
Titulek	<input type="text" value="Sbohem Simpsonovi!"/>	Rubrika	<input type="text" value="Život a vůbec"/>
Autor	<input type="text" value="Ivan Straka"/>	Postavení autora	<input type="text" value="redaktor"/>
Mluvčí	<input type="text"/>	Postavení mluvčího	<input type="text"/>
Kontext	<p>Na redakční poradě bylo odbojným redaktorům nařízeno rozkazem, že musím používat v souvislosti s myší sloveso "poklepat" (kliknout či cvaknout v žádném případě). Byla šance oživit českou terminologii, ať už fonetickým bastardem "kliknout" či ryze českým bufetáckým "cvaknout". Zvítězilo absurdní "poklepat", ač činnost, kterou s myší (polohovacím to zařízením) provádíme, má k poklepání tak daleko jako Mohamed k hoře.</p>		
Poznámka	<input type="text"/>		
Zdroj	<input type="text" value="Softwarové noviny"/>	Datum	<input type="text" value="16.10.2013"/>
Číslo	<input type="text" value="10"/>	Strana	<input type="text" value="132"/>
		Rok	<input type="text" value="1995"/>

 Tweet

Error

[Exportovat](#)

Dord

Ghost word

- dord
 - Webster's New International Dictionary 2nd ed. (1934)

DOR·COP'SIS (dôr·kôp'sis), *n.* [NL., fr. Gr. *dorkas* gazelle + *-opsis*.] *Zool.* A genus of small kangaroos of Papua.
dord (dôrd), *n.* *Physics & Chem.* Density.
||do're' (dô'rā'), *adj.* [F.] **a** Golden in color. **b Metal.** Containing gold: as. *doré* silver. — *n.* = DORÉ BULLION.

D or d, cont.
Physics & Chem.
density.

- 1939 objeveno, 1940 smazáno v knize

Evidence-based

Korpusy!

První slovník z korpusu: COBUILD.

Fáze tvorby slovníku

Analýza (dat). → Syntéza (slovníku).

Korpusy – vývoj

1. Jednodušší tvorba.
2. Rozsah, rychlost.
3. Nástroje pro analýzu.
4. Automatické nástroje.

Co všechno se dá automatizovat?

Sběr dat (do korp.).

Rozp. významu.

Anotace dat.

Vytvoření hesla.

Analýza dat.

Příklady, překlady.

Definice?

Jsou slovníky k ničemu?

Potřeby uživatele

Co slovo znamená?

Jaký je jeho ekvivalent v mém jazyce?

Jak se chová k ostatním slovům?

Jsou slovníky k ničemu?

1. Přetrvává potřeba kvalitních materiálů.
2. Jejich výroba je čím dál jednodušší.

Experiment

Pozorujte, kolikrát denně používáte internetové slovníky.

Typy slovníků

Podle obsahu

a) encyklopedické

b) **jazykové**

c) diachronní (historické, etymologické)

Jazykové → Podle času

a) synchronní

b) diachronní

Synchronní → Podle počtu jazyků

a) jednojazyčné

b) vícejazyčné (překladvé)

Jednojazyčné → Podle cíle

a) normativní (obv. školní)

b) deskriptivní (vědecké)

Deskriptivní → Podle uspořádání

a) alfabetické

b) jiné – např?

Podle uspořádání

a) alfabtické

b) jiné

(retrográdní, frekvenční, věcné/tezaurové)

Podle oblasti

a) celonárodní

b) dílčí – např?

Podle oblasti

a) celonárodní

b) dílčí:

s. autora/díla, dialektické, slangové,

termino–, onoma–, frazeo–

Podle média uložení

... ale to už znáte.

... a další.

(V tomto předmětu se bavíme primárně o...

... výkladových slovnících.)

Zásada

Netvoříme **slovník** určitého typu.

Tvoříme slovník pro určitý typ **uživatele**.

Příprava projektu slovník

Příprava projektu slovník

Zdroje dat

Příprava projektu slovník

Zdroje dat

Uživatelské profily

Příprava projektu slovník

Zdroje dat

Uživatelské profily

Software, style guide

Příprava projektu slovník

Rozpočet, časový plán

Příprava projektu slovník

Rozpočet, časový plán

Rozsah, obsah, struktura, vzhled

Příprava projektu slovník

Rozpočet, časový plán

Rozsah, obsah, struktura, vzhled

Uživatel a užití

Uživatel

Věk

Odbornost

Znalost j.

Zkušenosti se s.

Užití – příručka

Neznámá
slova

Pravopis

Křížovky,
scrabble

Další užití

Studium

Překlady

Uživatelské profily

~ studium potřeb zákazníka.

Uživatelské profily

**Dotazníky, agentury,
weby.**

Uživatelské profily

Dotazníky, agentury,
weby.

Akademické.

Uživatelské profily

Dotazníky, agentury,
weby.

Časté vyhledávání.

Akademické.

Uživatelské profily

Dotazníky, agentury,
weby.

Časté vyhledávání.

Akademické.

Zpětná vazba.

Style guide

Style guide

~ pravidla pro opakující se jevy.

Style guide – např.

Struktura hesla.

Style guide – např.

Struktura hesla.

Pořadí hesel.

Style guide – např.

Struktura hesla.

Pořadí hesel.

Formátování.

Style guide – např.

Struktura hesla.

Zobrazení.

Pořadí hesel.

Formátování.

Style guide – např.

Struktura hesla.

Zobrazení.

Pořadí hesel.

Podoba databáze.

Formátování.

Style guide – např.

Struktura hesla.

Zobrazení.

Pořadí hesel.

Podoba databáze.

Formátování.

Zkratky.

Style guide

Pravidla.

Pokyny pro zadávání.

Doporučení.

Pilotní fáze.

Obecné principy.

“[Philip] Gove claimed he saved 80 pages in the Third by using fewer commas.”

(S. Fatsis 2015,

o editoru *Merriam-Webster’s Third International*)

Přehled

Preskripce a evidence-based výzkum.

Přehled

Preskripce a evidence-based výzkum.

Typy slovníků odrážejí potřeby uživatele.

Přehled

Preskripce a evidence-based výzkum.

Typy slovníků odrážejí potřeby uživatele.

Příprava slovníku je (pokud možno)

složitější než samotná tvorba.

Úkol

- 1) Přihlaste se do Lexonomy.
- 2) Vytvořte nový slovník s několika hesly a následující heslovou strukturou:

Úkol: Heslová struktura

– lemma

– POS (přidejte možnosti)

– výslovnost (min. i max. 1)

– význam (min. 1) s příklady a prázdnou expresivitou